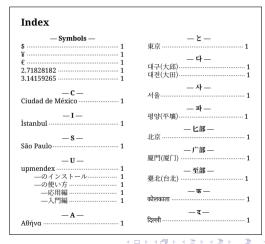
upmendexで作る多言語索引 Multilingual index processing by upmendex

田中 琢爾 TANAKA Takuji

2024年11月10日

Overview

- Feature of upmendex
 - Multilingual index processor
- Localization
 - Latin, Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, Thai
 - Arabic, Hebrew
- Multilingual environment
- Benchmark



Feature of upmendex (Ver. 1.10)

- Index processor
 - Upper compatible with MakeIndex/Mendex
 - Work with upLaTeX/LuaLaTeX/XeLaTeX
- Localization
 - Support 60 Languages / 12 Scripts
 - Latin (incl. non-English), Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, Thai
 - Arabic, Hebrew
 - Symbol, Number
- Multilingualization
 - Unicode, UTF-8
 - ICU[†] (Collation, Case Conversion, Category Property)
 - Environment for babel/polyglossia

† ICU: International Components for Unicode



Language, Script, Locale

Language	Script / Index	ICU locale			
English	Latin	root			
Spanish	Latin	es			
German	Latin	de			
Turkish	Latin	tr			
•••					
Russian	Cyrillic	ru			
Ukrainian	Cyrillic	uk			
•••	-				
Greek	Greek	el			
	Hanzi / Pinyin	zh			
Chinese	Hanzi / Stroke	zh-u-co-stroke			
Cilliese	Hanzi / Radical-Stroke	zh-u-co-unihan			
	Hanzi / Zhuyin	zh-u-co-zhuyin			
Japanese	Kana & Hanzi / Kana	ja			
Korean	Hangul	ko			

Language	Script / Index	ICU locale
Hindi	Devanagari	hi
Marathi	Devanagari	mr
•••		
Thai	Thai	th
Persian	Arabic	fa
Arabic	Arabic	ar
•••		
Hebrew	Hebrew	he
Yiddish	Hebrew	yi
Common	Symbol	
Common	Number	

upmendex supports 60 languages, 12 scripts & 95 locales.

Latin, Cyrillic, Greek

- Sorting (Collation) | ソート順
- Diacritical mark | ダイアクリティカルマーク
- Digraph/Trigraph | ダイグラフ/トライグラフ

German, Phonebook Sort Order

Collation Rule locale: de-u-co-phonebk

```
&AE<<ä<<Ä
&OE<<ö<<<Ö
&UE<<ü<<<Ü
```

Style File *.ist locale: de-u-co-phonebk

icu_locale "de-u-co-phonebk"

German Inputs in *.tex

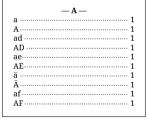
```
ad\index{ad}.
ae\index{ae}.
AE\index{AE}.
ä\index{ä}.
ä\index{ä}.
a\index{á}.
a\index{af}.
...
```

Index

```
-A-
a 1
A 1
ä 1
Ä 1
Ä 1
Ad 1
Ad 1
Ad 1
Ad 1
Ad 1
AD 1
Ae 1
Ae 1
AE 1
AF 1
```

default

Index



phonebook sort order

Lithuanian, Sort Order of Y

Collation Rule locale: It

Lithuanian Inputs in *.tex

```
i\index{i}.
I\index{I}.
i\index{i}.

i\index{i}.

y\index{y}.

y\index{y}.
```

Rodyklė _н_ h 1 $-\mathbf{x}$

Slovak, Diacritical Mark

Collation Rule locale: sk

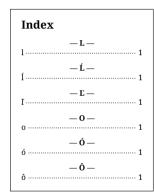
```
&0<ô<<<ô
```

Collation Rule locale: sk-u-co-search

```
&L<1<<<L<T'<<<L'
&O<ó<<<Ô<<<Ô
· · ·
```

Slovak Inputs in *.tex

default



General-Purpose Search

Turkish, Dotless / Dotted I

```
 \begin{array}{c|cccc} \textbf{language} & \textbf{upper} & \textbf{lower} \\ \hline \textbf{Turkish} & & \textbf{I} & \textbf{i} \\ \hline \textbf{English} & & \textbf{I} & \textbf{i} \\ \hline \end{array}
```

```
Collation Rule locale: tr
&[before 1]i<1<<<I
&i<<-İ
```

```
Turkish Inputs in *.tex

h\index{h}.

h\index{H}.

i\index{i}.

i\index{i}.

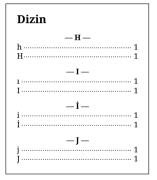
I\index{I}.

I\index{I},

j\index{j}.

J\index{J}.

...
```

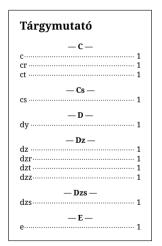


Turkish

Hungarian, Digraphs and Trigraph

Collation Rule locale: hu &C<cs<<<Cs &D<dz<<<DZ &DZ<dzs<<<DZS

Hungarian Inputs in *.tex cr\index{cr}. cs\index{cs}. ct\index{ct}. dy\index{dy}. dz\index{dz}. dzr\index{dzr}. dzs\index{dzs}.



. . .

dzt\index{dzt}.

e\index{e}.

Cyrillic & Greek

Russian Inputs in *.tex

цветок\index{цветок}. птица\index{птица}. ветер\index{ветер}. луна\index{луна}.

Greek Inputs in *.tex

 $λουλούδι\index{λουλούδι}.$ πουλί $index{πουλί}.$ άνεμος\index{άνεμος}. φεννάρι\index{φεννάρι}.

Предметный указатель

```
— B —
ветер ...... 1
      — л —
луна...... 1
      -\pi
птица ...... 1
      -II-
пветок------ 1
```

Russian, Cyrillic

Ευρετήριο

```
άνεμος...... 1
λουλούδι ...... 1
     -\pi
φεννάρι ...... 1
```

Greek

CJK (Chinese, Japanese, Korean)

- Chinese: 4 kinds of sort order | 中国語: 4種のソート順
- Japanese: Reading, Extended Kana

|日本語: 読み、仮名拡張

● Korean: composed/decomposed | 韓国語: 完成型・組合型

Chinese, Han Ideograph Sort Order

Style File *.ist locale: zh-u-co-unihan

```
%icu_locale "zh"
%icu_locale "zh-u-co-stroke"
icu_locale "zh-u-co-unihan"
%icu_locale "zh-u-co-zhuyin"
hanzi_head "一部;|部;\部;\J部;\Z部;\J部;二部;
```

Chinese Inputs in *.tex

花\index{花 (8, 艸, huā, 厂メ丫)} 鳥\index{鳥 (11, 鳥, niǎo, ろー幺ˇ)} 風\index{風 (9, 風, fēng, 匸L)} 月\index{月 (4, 月, yuè, 凵せ`)}

sort orde	locale	
Pinyin	拼音	zh
Stroke	筆畫數	zh-u-co-stroke
Radical-Stroke	部首筆畫數	zh-u-co-unihan
Zhuyin (Bopomofo)	注音符號	zh-u-co-zhuyin

Chinese, Han Ideograph Sort Order

Pinyin Sort Order 拼音 locale: zh

```
\centerline{\bfseries --- F ---}\par\nobreak
\item 風 (9, 風, fēng, □乙)\leaders\hbox{ }\hfill 1
\indexspace
\centerline{\bfseries --- H ---}\par\nobreak
\item 花 (8, 艸, huā, 厂メY)\leaders\hbox{ }\hfill 1
```

Radical-Stroke Sort Order 部首筆畫數 zh-u-co-unihan

```
\centerline{\bfseries --- 月部 ---}\par\nobreak \item 月 (4, 月, yuè, 凵世`)\leaders\hbox{ }\hfill 1 \indexspace \centerline{\bfseries --- 艸部 ---}\par\nobreak \item 花 (8, 艸, huā, 厂XY)\leaders\hbox{ }\hfill 1
```

Stroke Sort Order 筆畫數 zh-u-co-stroke

```
\centerline{\bfseries --- 四書 ---}\par\nobreak
\item 月 (4, 月, yuè, ロせ`)\leaders\hbox{ }\hfill 1
\indexspace
\centerline{\bfseries --- 八書 ---}\par\nobreak
```

\item 花 (8, 艸, huā, 厂XY)\leaders\hbox{ }\hfill 1

```
upmendex Output *.ind
```

Zhuyin (Bopomofo) Sort Order 注音符號 zh-u-co-zhuyin

```
\centerline{\bfseries --- □ ---}\par\nobreak
\item 風 (9, 風, fēng, □∠)\leaders\hbox{ }\hfill 1
\indexspace
```

\item 鳥 (11, 鳥, niǎo, ラー幺)\leaders\hbox{ }\hfill

\centerline{\bfseries --- \centerline{\par\nobreak

Chinese, Han Ideograph Sort Order

索引

索引

索引

索引

拼音 pinyin 筆畫數 stroke 部首筆畫數 radical-stroke 注音符號 zhuyin (bopomofo)

Chinese, Polyphone (多音字)

Input with Polyphone in *.tex

```
重新\index{重新 (chóng xīn)}
重要\index{重要 (zhòng yào)}
長年\index{長年 (cháng nián)}
長短\index{長短 (cháng duǎn)}
長大\index{長大 (zhǎng dà)}
```

Style File *.ist

icu_locale "zh" %icu_locale "zh-u-co-zhuyin" icu_rules "&常<<長年<<長短 &崇<<重新"

索引

```
      - C ー

      長短 (cháng duǎn)
      1

      長年 (cháng nián)
      1

      重新 (chóng xīn)
      1

      - Z ー
      長大 (zhǎng dà)
      1

      重要 (zhòng yào)
      1
```

拼音 pinyin

索引

注音符號 zhuyin (bopomofo)

Japanese, Sort by Reading (Yomi)

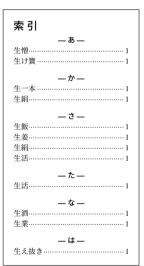
Japanese Inputs with Reading (Yomi) in *.tex

```
\newcommand{\YomiTag}[1]{\relax}

% \index{reading@index_word}
生酒\index{なまざけ@生酒}。
生一本\index{きいっぽん@生一本}。
生け簀\index{いけす@生け簀}。
生絹\index{きぎぬ@生絹\YomiTag{きぎぬ}}%
\index{すずし@生絹\YomiTag{すずし}}。
生飯\index{さば@生飯}。
生姜\index{しょうが@生姜}。
生活\index{せいかつ@生活\YomiTag{せいかつ}}
\index{たつき@生活\YomiTag{たつき}}。
```

Japanese words consist of Hanzi (ideographs) & Kana (syllabaries), sorted by reading (yomi), indexed by Kana.

This feature is implemented by ASCII mendex.



Japanese, Reading & Dictionary

Japanese Inputs in *.tex

```
生酒\index{生酒}。
生一本\index{生一本}。
生け簣\index{生け簣}。
生絹\index{生絹}%
\index{生絹}%
とえ抜き\index{生紀}5。
```

Dictionary *.dic

```
index_word reading
生酒 なまざけ
生一本 きいっぽん
生け簀 いけす
生絹 きぎぬ
生え抜き はえぬき
```

```
索引
生まれ付き ......1
    生さぬ仲.....1
生い立ち......1 生酒......1
生一本 ...... 1
```

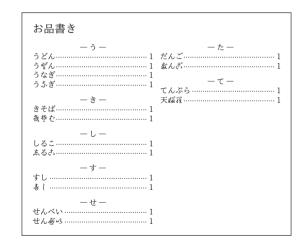
Japanese, Hentaigana

Inputs with Hentaigana in *.tex

```
後をで\index{後をで}
うなぎ\index{うなぎ}
えるよ\index{太るよ}
まし\index{まし}
天城程\index{天城程}
```

Dictionary for Hentaigana *.dic

```
index_word reading
き
き
そ
は
な
し
こ
```

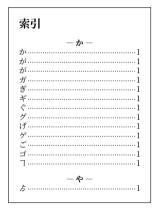


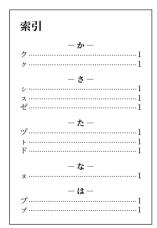
Extended Kana in JIS X 0213

Extended Kana Inputs in *.tex

Aynu itak Inputs in *.tex

```
ク\index{ク}
シ\index{シ}
ス\index{ス}
プ\index{プ}
せ\index{ゼ}
ヅ\index{ヅ}
ド\index{ド}
```





Japanese, Archaic Kana

Archaic Kana Inputs in *.tex

12\index{12}. % Hiragana Archaic YE % or Hentaigana E-1

[于\index{[于}. % Hiragana WU レ\index{レ}. % Katakana YI エ\index{エ}. % Katakana YE 于\index{于}. % Katakana WU

...

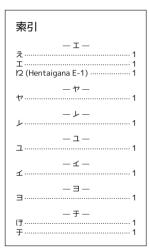
Dictionary: "" is Hentaigana E-1

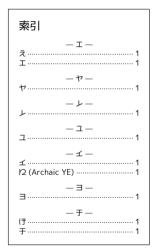
12

え

Style: "12" is Hiragana Archaic YE

icu_rules "&ゆ<口<<<エ<よ"





Korean, Modarn / Archaic Hangul

ideograph ^{한자}





	Unicode block	style		upmendex	upLaTeX	XeLaTeX
modarn	Hangul Syllables	composed	완성형	✓	√	
ex. 일	Hangul Jamo	decompopsed	조합형	✓	N.A.	✓
archaic	Private Use Area	composed	완성형	via dictionary	√	
ex. ·싏	Hangul Jamo	decomposed	조합형	✓	N.A.	\checkmark

Korean Hangul

Hangul Inputs in *.tex

```
M\index{M (composed)}.

M-\index{M- (decomposed)}.

M \index{M (archaic)}.

M :\index{M : (archaic with tone mark)}.

M\index{M (Hanyang PUA)}.
```

Dictionary for PUA code *.dic

```
Hanyang PUA code decomposed

사
씨 사 . I
```

찾아보기 스 (composed)1 소 (archaic)1 쓰 (composed)1 쓰 (decomposed)1 :☆ (archaic with tone mark).....1 쏘 (archaic)1 쌋 (Hanyang PUA)] △ (archaic)1 △ (archaic)1

Complex Text Layout

- Devanagari, Thai
- Arabic, Hebrew: R-to-L typeset
- Symbol, Number

Devanagari & Thai (experimental)

Hindi Inputs in *.tex

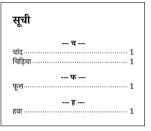
```
फूल \index{ फूल }
चिड़िया \index{ चिड़िया }
हवा \index{ हवा }
```

चांद \index{ चांद }

...

Thai Inputs in *.tex

```
ดอกไม้ \index{ ดอกไม้ }นก \index{ นก }ลม \index{ ลม }ดวงจันทร์ \index{ ดวงจันทร์ }
```



Hindi, Devanagari



Thai

Typeset by XeLaTeX

Arabic & Hebrew (experimental)

Arabic Inputs in *.tex

```
{زهرة \index{زهرة
{ عصفور }\index\ عصفور
{ریح }index{ریح
{القمر }index\القمر
...
```

Hebrew Inputs in *.tex

{פֶּרַח }\index ציפור \index{ ציפור } רוחַ \index{ רוּחַ } ירח \index{ ירח }





Arabic

Hebrew

R-to-L typeset by XeLaTeX. **upmendex** processes only indexing.

...

Symbol, Number

Script	charType	example	treatment by upmendex
Latin	Lu, Ll, Lo, : letters	ABCabc A a ⓐ	directly pass to ICU collator
Greek	Lu, Ll, Lo, : letters	ΑΒΓαβγ	direct
Cyrillic	Lu, Ll, Lo,: letters	АБВабв	direct
Kana	Lo: other letter	あいうアア⑦ニロ゚	direct
Hangul	Lo: other letter	가나다"ㄴ마쏎	direct
Hanzi	Lo: other letter	花鳥風月	lookup dictionary or direct
_	Lm: modifier letter	-,	direct
Number	Nd : dicimal digit number	012 0 1 2	direct
Number	No: other number	¹² 34 5 6 7,(8)9.	lookup dictionary or direct
	Sk: modifier symbol	, * * * * * * * * * * * * * * * * * * *	direct
	Sm: math symbol	÷⊳#	lookup dictionary or direct
Symbol	So: other symbol	ﯛ₫☎♥❸◎	lookup dictionary or direct
	Sc: currency symbol	€\$\$¢£¥¥₩₩	lookup dictionary or direct
	Po, Pd, Mn, Me,: other punctuation etc.	?!!?¡; †#§¶—	direct
_	Cc : control character	ESC, BS, DEL	ignore
_	Cf : format character	BOM, RLM	diect
others	Lo, etc. (unknown scripts)		lookup dic or direct (option "-f") or ignore

Characters are classified by Unicode *General Category Values* or "charTypes"

Ref. https://unicode.org/reports/tr44/#General_Category_Values



Multilingual Environment with upLaTeX/pxbabel

Block Setting for Scripts in Style File *.ist

```
script preamble
                  cvrillic "\n\\fontencoding{T2A}\\selectfont"
                  cvrillic "\n\\fontencoding{T1}\\selectfont"
script postamble
script_preamble
                  hangul
                           "\n\\begin{otherlanguage}{korean}"
                           "\n\\end{otherlanguage}"
script postamble
                  hangul
script_preamble
                  hanzi
                           "\n\\begin{otherlanguage}{tchinese}"
script_postamble
                           "\n\\end{otherlanguage}"
                  hanzi
```

upmendex Output *.ind

```
\centerline{\bfseries --- C ---}\par\nobreak
  \item Coфия\leaders\hbox{~}\hfill 1
\fontencoding{T1}\selectfont
  \indexspace
\centerline{--- サ ---}\par\nobreak
  \item さいたま\leaders\hbox{~}\hfill {1}
  \item 札幌\leaders\hbox{~}\hfill {1}
\begin{otherlanguage}{korean}
  \indexspace
\centerline{\bfseries --- □ ---}\par\nobreak
  \item 부산(釜山)\leaders\hbox{~}\hfill 1
\end{otherlanguage}
```

Multilingual Environment with XeLaTeX/polyglossia

Block Setting for Scripts in Style File *.ist

```
cyrillic "\n\\begin{russian}"
script_preamble
script_postamble
                  cyrillic "\n\\end{russian}"
script preamble
                           "\n\\begin{iapanese}"
                  kana
                           "\n\\end{iapanese}"
script postamble
                  kana
script_preamble
                  hangul
                           "\n\n\\begin{korean}"
                           "\n\\end{korean}"
script_postamble
                  hangul
script_preamble
                  hebrew
                           "\n\\begin{hebrew}"
script postamble
                           "\n\\end{hebrew}"
                  hebrew
```

upmendex Output *.ind

```
\centerline{--- さ ---}\par\nobreak
  \item さいたま\leaders\hbox{~}\hfill {1}
\end{ iapanese}
\begin{korean}
  \indexspace
\centerline{--- □ ---}\par\nobreak
  \item 대구(大邱)\leaders\hbox{~}\hfill {1}
\end{korean}
\begin{hebrew}
 \indexspace
\centerline{--- א ---}\par\nobreak
\item אשדוד \leaders\hbox{~}\hfill {2}
\end{hebrew}
```

Output of Multilingual Index

索 引	
— symbols —	— v —
€1	さいたま1
$3.14159265 \cdots 1$	札幌1
— c —	— F —
Ciudad de México······1	東京1
-1-	— E —
·	
Istanbul1	대구 (大邱)1
— s —	대전 (大田)1
São Paulo1	— A —
Sao I auto	서울1
- 6 -	\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
Београд2	— m —
Бишкек2	평양 (平壤)1
— к —	— 五畫 —
Київ2	北京1
— w —	— +三書 —
Москва2	廈門 (厦门)1
— ** —	— 十四書 —
大阪1	
7304	250 (1470)

with upLaTeX & pxbabel

Index	
— Symbols —	- 파 - 평양(平壤)
€	평양(平壤)
3.14159265 1	— ヒ部 —
— I —	北京
İstanbul 1	— 至部 —
-s-	臺北(台北)
São Paulo 1	— द —
— A —	दिल्ली
— A — Αθήνα ······ 2	— ग —
·	मुंबई
-e-	
Θεσσαλονίκη2	— ก — กรุงเทพมหานคร
— K —	•
Київ 1	— u —
— M —	นนทบุรี
Москва1	-1-
	ظبي
— さ — さいたま 1	— s —
	2
- Ł -	_,_
東京 1	שלים
_ 사 _	—n—

with XeLaTeX & polyglossia

Benchmark

	makeindex	mendex	upmendex	xindy
internal encoding	8bit 1byte	EUC-JP	UTF-16	Unicode
Collator	locale	ASCII, Kana	ICU collator	
Latin	√	√ ASCII	Lang:37, Locale:62	Lang:32
Greek			Lang:1, Locale:1	Lang:1
Cyrillic			Lang:9, Locale:9	Lang:6
Chinese			Lāng:1, Locale:4	
Japanese		✓ (Yomi, Dict)	√ (Yomi, Dict)	
Japanese		(Tollii, Dict)	Lang:1, Locale:2	
Korean			Lang:1, Locale:3	
¯ Devanagari			Lang:3, Locale:3	
Thai			Lang:1, Locale:1	
Arabic			Lang:6, Locale:7	
Hebrew			Lang:2, Locale:3	Lang:1
Ōther				Lang:4
Total		Lang:2	Lang:60, Locale:95	Lang:44



Languages by Number of Native Speakers

	Language	Script	speakers	ICU	polyglossia	upmendex	xindy
1	Chinese	Hanzi	1,370,000,000	√	√	√	
2	Engish	Latin	530,000,000	✓	✓	✓	\checkmark
3	Hindi	Devanagari	490,000,000	✓	✓	✓	
4	Spanish	Latin	420,000,000	✓	✓	✓	✓
5	Arabic	Arabic	230,000,000	✓	✓	✓	
6	Bengali	Bengali	220,000,000	V	√		
7	Portuguese	Latin	215,000,000	✓	✓	✓	✓
8	Russian	Cyrillic	180,000,000	✓	✓	✓	✓
9	Japanese	Kana & Hanzi	134,000,000	✓	✓	✓	
10	German	Latin	130,000,000	✓	✓	✓	\checkmark
11	French	Latin	123,000,000		√	√	$\overline{}$
12	Punjabi	Gurmukhi	90,000,000	✓	✓		
13	Javanese	Latin	75,000,000	✓	✓	✓	
14	Korean	Hangul	75,000,000	✓	✓	✓	
15	Vietnamese	Latin	70,000,000	✓	✓	✓	✓
16	Telugu	Telugu	70,000,000	✓	√		
17	Marathi	Devanagari	68,000,000	✓	✓	✓	
18	Tamil	Tamil	74,000,000	✓	✓		
19	Persian	Arabic	46,000,000	✓	\checkmark	\checkmark	
20	Urdu	Arabic	61,000,000	✓	✓	✓	

Ref. https://ja.wikipedia.org/wiki/ネイティブスピーカーの数が多い言語の一覧



Languages used on the Internet

	Language	share	Script	ICU	pg	upm	xnd		Language	share	Script	ICU	pg	upm	xnd
1	English	63.4 %	Latin	root	√	√	$\overline{}$	21	Ukrainian	0.3 %	Cyrillic	uk	V	$\overline{}$	$\overline{}$
2	Russian	7.1 %	Cyrillic	ru	\	/	\	22	Hebrew	0.3 %	Hebrew	he	\checkmark	✓	\checkmark
3	Spanish	3.9 %	Ĺatin	es	\	/	\	23	Swedish	0.3 %	Latin	sv	\checkmark	✓	✓
4	German	3.7 %	Latin	de	/	/	/	24	Romanian	0.3 %	Latin	ro	\	\	✓
5	Turkish	3.5 %	Latin	tr	/	\	\	25	Hungarian	0.3 %	Latin	hu	\checkmark	✓	✓
6	Persian	2.5 %	Arabic	fa	\	√		26	Danish	0.2 %	Latin	da	\	√	· · · · · · · · · · · · · · · · · · ·
7	French	2.0 %	Latin	root	✓	✓	✓	27	Slovak	0.2 %	Latin	sk	\checkmark	✓	\checkmark
8	Japanese	1.9 %	Kana	ja	\checkmark	✓		28	Serbian	0.2 %	Latn, Cyrl	sr	\checkmark	✓	✓
9	Portuguese	1.8 %	Latin	pt	\checkmark	✓	✓	29	Bulgarian	0.1 %	Cyrillic	bg	\checkmark	✓	\checkmark
10	Chinese	1.3 %	Hanzi	zh	\checkmark	✓		30	Finnish	0.1 %	Latin	fi	\checkmark	✓	✓
11	Vietnamese	1.3 %	Latin	vi	√	√		31	Croatian	0.1 %	Latin	hr	\checkmark	$\overline{}$	$\overline{}$
12	Italian	1.0 %	Latin	root	\checkmark	✓	\checkmark	32	Lithuanian	0.1 %	Latin	lt	\checkmark	✓	\checkmark
13	Arabic	0.9 %	Arabic	ar	\checkmark	\checkmark		33	Norwegian (Bokmål)	0.1 %	Latin	nb	\checkmark	✓	✓
14	Polish	0.9 %	Latin	pl	\checkmark	\checkmark	\checkmark	34	Hindi	0.1 %	Devanagari	hi	\checkmark	✓	
15	Greek	0.7 %	Greek	el	\checkmark	\checkmark	✓	35	Norwegian (nynorsk)	0.1 %	Latin	nn	\checkmark	\checkmark	\checkmark
16	Dutch	0.7 %	Latin	nl	\	√	√	36	Slovenian	0.1 %	Latin	sl	\	√	· · · · · · · · · · · · · · · · · · ·
17	Indonesian	0.7 %	Latin	root	\checkmark	\checkmark		37	Latvian	0.1 %	Latin	lv	\checkmark	✓	\checkmark
18	Korean	0.6 %	Hangul	ko	\checkmark	\checkmark		38	Estonian	0.1 %	Latin	et	\checkmark	\checkmark	\checkmark
19	Czech	0.4 %	Latin	cs	\checkmark	\checkmark	\checkmark	39	Azerbaijani	< 0.1 %	Latin	az		✓	
20	Thai	0.4 %	Thai	th	\checkmark	✓		40	Catalán	< 0.1 %	Latin	root	\checkmark	\checkmark	

Ref. https://ja.wikipedia.org/wiki/インターネットにおける言語の使用



To Do

- Support more scripts
 - Bengali, Telugu, Tamil, Malayalam, Kannada, Gujarati, Oriya, Sinhala
 - Khmer, Lao, Myanmar (Burmese)
 - Tibetan, Mongolian
 - Armenian, Georgian
 - Ethiopic (Amharic)
 - etc.
- Support more locales
 - Latin Script: Sorbian, Hausa, Igbo, Yoruba, Kalaallisut, Breton, Uzbek
 - etc.
- Add style options
 - script_head

Feedback is welcome

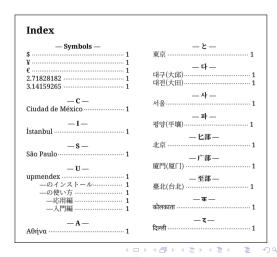
https://github.com/t-tk/upmendex-package/issues



Summary

I introduced multilingual index processor **upmendex**.

- Feature
- Localization: 60 Languages, 12 Scripts
 - Latin, Cyrillic, Greek
 - CJK (Chinese, Japanese, Korean)
 - Devanagari, Thai, Arabic, Hebrew
 - Symbol, Number
- Multilingualization
 - Environment for upLaTeX/babel & XeLaTeX/polyglossia



References

- ASCII Nihongo TeX (Publishing TeX), ASCII MEDIA WORKS (web site by DWANGO Co., Ltd.). The site distributes mendexk source files.
- Source/Document distribution of upmendex multilingual index processor @ GitHub. upmendex @ CTAN
- upTeX, upLaTeX unicode version of pTeX, pLaTeX
- International Components for Unicode (ICU)
- PXbase LaTeX: Support library for other PX packages @ GitHub. The repository distributes pxbabel. pxbase @ CTAN
- opolyglossia An alternative to Babel for XeLaTeX and LuaLaTeX @ GitHub. polyglossia @ CTAN
- "Indexing Makes Your Book Perfect" by SHIKANO Keiichiro at TUG2013, October, Tokyo.

